面向情报研究的文本语义挖掘方法述评*

赵冬晓 王效岳 白如江 刘自强
(山东理工大学科技信息研究所 淄博 255049)

摘要: 【目的】对主要的文本语义挖掘方法及其在情报研究中的应用进行综述分析。【文献范围】集中选择近 10 年国内外主流的文本语义挖掘方法在情报研究领域的应用以及少数此前的代表性研究和文本语义挖掘方法的进展研究。【方法】分别概括介绍词、句子和篇章粒度的文本语义挖掘方法、算法，并通过主题演化和技术挖掘领域的实际应用进行方法剖析。【结果】文本语义挖掘方法与传统的情报分析方法相比，主要弥补了两个缺陷：侧重于分析结构化的数据，无法处理多种异构的数据源；分析停留在统计语法层面，没有深入到文本的语义信息。【局限】仅对主流的文本语义挖掘方法以及在科学研究领域的应用进行综述分析，研究不全面。【结论】文本语义挖掘方法弥补了传统情报分析方法的不足，是情报研究方法的重要发展方向，随着方法的成熟，下一步研究重点是外部语义资源的丰富。

关键词: 文本语义挖掘 情报分析 主题演化 技术挖掘
分类号: G350

1 引 言

21世纪伊始，网络信息及通讯技术的发展造成电子信息爆炸，数据量每天以EB的单位增长，IDC和EMC联合发布的“2020年的数字宇宙”报告预测到2020年，全球数字宇宙将会膨胀到4000EB，根据调查结果和服务器容量调查可以得到一个相对合理的推断：目前，全球产生的数据量中仅有1%左右的数据能够被保存下来，也就是说今天全球能够被保存下来的数据在50EB左右，而其中被标记并用于分析的数据更是不到10%，这些信息数量巨大，无统一结构，难以被人或计算机所利用，但是蕴含着巨大的价值。

情报资源数据量急剧增加，如何利用这些结构多样化的信息，并从中准确快速地挖掘出有价值的情报成为情报工作者面临的难题。情报工作面临以下两个方面的挑战:

(1) 传统的情报分析方法与多种异构的数据源之间的矛盾。

传统的情报分析方法更多的是对文献的题录信息和引文信息以及其他结构化数据源进行分析，因此面对当今结构多样化的信息资源就显得无能为力，大大限制了情报的获取渠道。

(2) 传统的情报分析方法与文本内容的深层次挖掘要求之间的矛盾。

传统的分析方法一般停留在简单的统计语法分析层面，没有深入到文本内容的语义层面，造成语义的缺失问题，这样分析出来的情报不够准确和完整，也影响了知识整合。

随着数据挖掘、自然语言处理技术的发展和成熟，文本语义的深度挖掘成为可能，文本语义挖掘可以帮助情报人员进行准确的情报挖掘和分析工作，在信息处理、科学研究等领域有着广阔的应用前景。

本文通过调研国内外文本语义挖掘的研究现状，归纳了文本语义挖掘的主要技术方法，并对其在科研领域的应用进行详解，以期为文本语义挖掘在情报分析工作中的应用提供帮助。

通讯作者：赵冬晓，ORCID：0000-0002-9518-4281，E-mail：927011467@qq.com。
*本文系国家社会科学基金一般项目“未来新兴科学研究前沿识别研究”(项目编号：16BTQ083)的研究成果之一。

2 文本语义挖掘的主要方法

文本语义挖掘[1]是在文本挖掘的基础上寻找文中的语义模式，进行文本语义分析的过程，文本语义挖掘按照处理粒度分为词、句子和篇章三个级别，低粒度的文本语义挖掘方法既可以独立承担部分情报分析任务，如目标信息的抽取，优化共词分析等，同时还是高粒度的文本语义挖掘的基础，以便从宏观上进行情报分析，如主题演化分析以及研究前沿探测等。

2.1 词粒度语义挖掘

词粒度的语义挖掘是文本语义挖掘的基础工作，由于该粒度的处理单位是词语，因此该方法不受数据结构和格式的约束，只要是文本信息都可以在该粒度进行语义挖掘；该粒度的语义挖掘方法主要有词性标注和词义消歧，通过词粒度的语义处理，可以为文本数据附加第一层语义信息，并且为句子和篇章级别进行的文本语义挖掘奠定了基础。

(1) 词性标注(POS Tagging)

词性标注(Part Of Speech Tagging或POS Tagging)是指对句子中的每一个词都指派一个词性，如名词、动词或形容词，又称词类标注或者简称标注，目前的词性标注算法大体分为[2]：基于规则的方法、基于统计的方法和规则与统计相结合的方法。

最初的词性标注系统是采用基于规则的方法，著名的TAGGIT系统利用3 300条上下文规则，对100万词的Brown语料库进行标注，准确率达到了77%[3]。但是如果针对某一种语言的各种语言现象都构造规则的话，是非常的艰难和耗时的，而且根据规则判断词性的时候面临多种选择，脱离上下文很难做出正确的选择。

基于统计方法是20世纪80年代初随着统计学在计算语言学中的重新崛起兴起的，也是现在最常用的一种方法，常见的有基于N元模型的方法和基于隐马尔可夫模型的方法，近年来决策树、最大熵[4]、条件随机场[5]和SVM[2]等也被用于词性标注，并取得了不错的效果。

规则和统计相结合的方法弥补了两种方法的不足，国内北京大学计算语言学研究所提出了一种先规则后统计的规则和统计相结合的算法，正确率达到96.6%[6]。

现在的自然语言处理工具基本都具有词性标注的功能，中文主要有中国科学院计算技术研究所的ICTCLAS[7]和 LTP 语言技术平台[8]，英文的词性标注器有 Stanford Log- linear Part-Of-Speech Tagger[9]以及 CLAWS POS Tagger[10]等，另外还有一些开源的工具包[11]也提供词性标注支持。

经过词性标注，每一个词都有了例如：IN 介词、CD 数词、JJ 形容词性、NNP 专有名词、NN 名词等计算机可识别的类别标签，在此基础上可以快速准确地识别出文本中情报含量最多的信息，如数词、动词和名词等具有实际意义的词，实现各种格式的全文本语义挖掘。文献[12]在词性标注的基础上，将名词、动词、形容词、副词等具有较强的语义内容的词构成微博主题词，并进一步进行词性过滤和停用词处理，优化了共词网络，实现了基于共词网络的微博文本特征提取，其结果优于基于文档频率的方法，取得了更好的话题的识别效果。文献[13]针对电子商务领域的在线评论，提出了一种情感标签抽取方法，即识别产品特征和评价词之间是否存在修饰关系，该方法在词性标注的基础上实现对名词性信息和形容词性信息的抽取，并通过最大熵进行情感标签过滤，最终得到情感标签的集合。文献[14]在文本情感计算研究中，创新性地在依存句法分析结果的基础上，对句子进行情感主干抽取，根据依存关系的不同和词性搭配的不同定义了情感计算规则，进行情感倾向性分析，有效提高了情感分析的准确性。文献[15]提出利用文本语义挖掘构建中文领域本体的方法，通过词性标注、依存句法分析以及模式匹配等方法，从非结构化的文本中自动抽取术语和关系，实验结果表明该方法构建的本体能更好地反映领域知识结构。

综上所述，词性标注作为一种最基础的文本语义挖掘方法，既可以单独承担简单的信息抽取的功能，同时又可以作为辅助，对基本的情报分析方法进行语义改善，优化情报分析结果。

(2) 词义消歧 (Word-sense Disambiguation)

词义消歧是在词语具有多个意思的时候，识别出词语在特定句子中的具体词意。作为自然语言处理的底层研究，词义消歧早在 20 世纪 40 年代早期就已成为机器翻译中的一个确定的难题，Weaver 在 1949 年论及机器翻译时也肯定了词义消歧的重要意义[16]。

表 1 按照所用语义资源的不同，将词义消歧分为

基于知识库和基于语料库两种方法，在基于语料库的方法中，又可以根据是否有人工干预分为有监督的消歧和无监督的消歧。

表 1　词义消歧方法

分类	常用方法	优点	缺点
基于知识库	基于词典资源；基于知识本体；固定搭配等	不需要训练语料也不需要对词典资源进行人工处理；能进行大规模的词义消歧	知识缺乏完备性
基于语料库	有监督的词义消歧：机器学习算法(决策树、支持向量机、最大熵)	消歧正确率高	难以应用于大规模词义消歧
	无监督的词义消歧：聚类算法	无需人工干预	仅能区分词义类别，无法对词义进行明确标注

在基于知识库的词义消歧中，基于词典资源是最常用的方法，基于词典的语义消歧始于1986年[17]，Lesk[18]直接利用词典的词义解释或者定义来指导歧义词的词义判断，但是正确率在50%-70%之间，不是很理想。1988年, Pook和Catlett[19]提出另外一种改进方法，对上下文的词语进行同义词扩展，可以增大计算覆盖度的成功率。1995年，Agirre等[20]采用WordNet的分类体系计算歧义词及其上下文词语的概念密度，正确率达到80%左右。但是词典资源面向的词义消歧大都是通用文本，在面对特定领域的词义消歧时，由于特定的上下文环境和特定的词义变化使得基于词典资源的方法无法取得好的效果，因此基于领域本体资源的词义消歧成为发展趋势[21]。

基于语料库的词义消歧中，机器学习算法是最常用的有监督的词义消歧方法，有监督的词义消歧常被看做是分类问题，常用的算法有决策树、最大熵[22]、向量空间模型[23]等；无监督的词义消歧往往被看做是词语聚类问题[24]。

BioNLP在2011年的共享任务中实现了词义共指消歧，目前该功能可以通过软件[25]使用，识别出所有指代同一对象的词或词组，实现语义泛化，避免产生歧义。文献[26]为了对突发事件进行结构化信息抽取，提出了描述突发事件案例的四元特征向量模型并基于这一模型构建抽取框架，在事件时间信息、地点信息和其他侧面信息的抽取过程中，都运用指代消解对时间、地名和候选信息进行处理，在突发事件的要素抽取上达到了较高的准确率和召回率。文献[27]为了从学术期刊中抽取其中的理论，将理论识别视为命名实体识别问题，提出基于语义泛化思想的命名实体识别方法，实验选择词性标记和中国知网的义原作为泛化的方法，选择依据词性标记选择义项的词义消歧的方法，采用CRF模型进行实验，达到了较高的识别准确率。

词义消歧作为一种语义改善方法，能够有效地解决数据稀疏高维的问题，提高情报分析结果的准确性和完善性，目前在各类信息抽取任务中得到广泛应用。

2.2　句子粒度语义挖掘

句子粒度的文本语义挖掘方法主要是语义角色标注，该方法适用于任何具备句法结构的信息；由于兼具词粒度的语义信息和句法结构信息，比词粒度的文本语义挖掘更加完整，成为文本语义挖掘的关键技术。

(1) 语义角色标注(Semantic Role Labeling, SRL)

语义角色标注是在句子级别进行浅层的语义分析，标注句子中某些短语为给定谓词的论元 (语义角色)，如施事、受事、时间和地点等。

语义角色标注方法主要有[28]：基于句法分析的语义角色标注、基于特征向量的语义角色标注、基于机器学习的语义角色标注。

基于句法分析的语义角色标注又分为基于短语结构、基于组块分析和基于依存句法的方法，基于短语结构句法分析技术比较成熟，结果比较稳定，但是存在语料的稀疏严重、难以抽取更有效的特征等问题，因此很难再使结果有进一步的提高，越来越多的研究开始转向基于依存树的语义角色标注方法。Hacioglu等[29]首次基于依存句法实现了英文的语义角色标注，所使用的依存树是由句法树转换而来，并使用SVM分类器实现了角色分类，提出12个特征(依存关系、位置、中心词等)。文献[30]采用英文的基于依存关系的语义角色标注方法，实现了基于中文依存句法分析的语义角色标注系统。

基于特征向量的语义角色标注就是从句子中抽取所有充当语义角色的句法成分，然后进行角色的识别和分类，并进行角色标注。Gildea等[31]首先提出基于短语结构句法分析 SRL系统的7个基本特征(谓词原型、谓词词性、子类框架、位置、路径、依存关系、中心

词); Pradhan等[32]在基本特征基础上又引入了命名实体、中心词词性、谓词类别、部分路径等12种新特征; 李世奇等[33]提出一种基于特征组合和支持向量机的语义角色标注, 该方法以句法成分作为基本标注单元, 首先从当前基于句法分析的语义角色标注系统中选出高效特征, 构成基本特征集合, 然后提出一种基于统计的特征组合方法, 语义角色标注整体F值达到91.81%, 提高了近2%。王红玲[34]对基于特征向量的语义角色标注进行了系统而深入的研究与探索。

基于机器学习的语义角色标注能够克服基于规则的语义角色标注方法依赖知识库的问题, 目前核函数、最大熵以及条件随机场[35-36]等都成功地应用于语义角色标注任务中。

目前具有语义角色标注功能的软件已经比较成熟, 例如用哈尔滨工业大学语言云[8]对句子"2013 年 3 月 5 日, 第 12 届全国人民代表大会第一次会议在北京召开。"进行语义角色标注后可以直接得到: "全国人民代表大会"表示动作的实事; "2013 年 3 月 5 日"是时间, "北京"是地名, 语义角色标注后的词或短语有了附加的属性, 使得计算机对语句有了"浅层"的语义理解。文献[37]提出一种利用语义分析技术识别科技文献的创新内容的方法, 该方法主要是以句子为最小的标引粒度, 通过 KeyGraph 算法抽取摘要中的关键词, 并与 WordNet 进行映射识别出语义角色, 据此进行特征选择, 用 SVM 对科技文献进行语义角色标注, 实验表明该方法能有效识别出科技创新的内容, 大大缩短了科技人员翻阅文献的时间。文献[38]借助领域本体, 在对句子进行语义角色标注的基础上, 结合句法分析对创新点句中的主题词及主题词对应属性实例进行识别, 进一步挖掘创新点句中的知识关系。文献[39]针对传统的科技创新主题概率识别方法忽略文本的语义内容的问题, 提出基于 LDA 的科技创新主题语义识别模型, 该方法在对科技文献进行语义角色标注的基础上, 构建LDA 主题语义识别模型, 根据表征科技创新内容的关键词语义角色对应的上位词的概率更加准确地识别出科技创新主题。

语义角色标注是一种简单灵活的语义挖掘方法, 其操作方法和粒度对于情报分析工作来说都是可接受的, 语义角色标注不仅适用于上述科技创新识别目的的情报研究工作, 同样适用于其他目的性较强的研究, 是在强调文本语义内容理解的前提下进行情报研究的最为简单直接的手段。

2.3 篇章粒度语义挖掘

篇章粒度的文本语义挖掘在一定程度上克服了传统的情报研究方法受限于格式化信息, 如题录信息分析的问题, 其主要贡献在于能够深入到文本内容的层面, 结合词粒度和句子粒度已经赋予文本的语义信息进行整体的文本语义挖掘, 从宏观角度揭示科技发展趋势和发展的大方向, 是主题演化和技术挖掘中的主要方法之一。

(1) 文本聚类

文本聚类是一种无监督的机器学习方法, 能够在没有给定类别的前提下根据信息内容相似度进行聚集, 快速、高质量地将大量信息组织成少数有意义的簇, 从而获取这些信息中隐藏的知识或模式, 在数据量极大的情况下能够帮助情报工作人员简单快速地把握大致信息, 为进一步深度情报分析奠定基础。

文本表示的方法和聚类算法对聚类效果有直接的影响, 表 2 总结了当前各环节使用的主要方法或算法。

向量空间模型是最为常用的文本表示方法, 但是由于其忽略了文本内容的语义关系, 之后出现了通过本体进行语义改善, 构建语义向量空间模型的方法[40]和 LSI 模型[41]来发现潜在语义关联的文本表示方法, 这些文本表示方法的改进都不同程度优化了文本的聚类效果。文本表示除了将文本表示成计算机可识别的形式以外, 更为深入的文本表示是文本的主题表示, 也就是主题挖掘, 目前对篇章粒度直接进行主题识别的方法主要是通过 LDA 模型。

LDA 模型也称为三层贝叶斯概率模型, 包含词、主题、文档三层结构, LDA 将每个文档表示为多个主题混合, 每个主题是多个主题词的混合, 主题服从主题词表上的一个多项式分布, 这些主题被数据集中的所有文档共享, 每个文档有一个特定的主题混合比例[42]。LDA 主题模型体现了主题、主题词和文档的三层语义结构, 使文本的主题具有更充分的语义信息和可解读性, 成为文本主题挖掘的主要方法之一。

由于文本聚类能够将大量的信息聚集成少数有意义的话题, 因此该方法在信息检索和管理领域得到广泛的应用。文献[40]提出了适用于知识库的树状结构的多层次聚类, 在领域本体的帮助下, 实现将词映射

表 2 文本聚类方法

文本表示方法	分析	相似度计算方法
向量空间模型	文档空间被看作一组由正交向量张成的向量空间，每一篇文档都被映射成多维向量空间中的一个点，用此空间中的向量来表示。向量空间模型的文本表示方法简单，具有良好的文本表示效果，但是由于其采用独立性假设，割裂了文本原有的语义关系，在对文本主题挖掘的准确性要求较高时，需要进一步进行语义丰富。	(1) 基于向量空间模型的相似度计算：余弦相似度；距离相似度(欧氏距离、幂距离、街区距离等)； (2) Jaccard 系数等； (3) 基于本体的相似度计算：文本之间的相似度被转换成概念之间的相似度。
语义向量空间模型	在文本表示中引入语义向量，其主要方法是通过本体实现文本中的词汇和本体概念之间的映射，从而实现词汇语义的丰富，该方法是对传统的向量空间模型的语义优化改进。	
LSI 模型	该模型将文本表示为词–文档矩阵，其核心思想在于通过奇异值分解将词向量和文档向量投影到低维的语义空间，一方面消减了原矩阵中的“噪音”，突出了词和文本的语义关系，另一方面能捕获到词之间的相关性，发现潜在的语义关联。	
后缀树模型	将文档看成一个由若干短语组成的字符串，一个短语就是具有一个或者更多个词的有序序列，该方法优于向量空间表示中词语之间互相独立而导致的语义缺失问题，在英文文档聚类中使用较多。	基于短语的相似度计算：主要用于后缀树模型中文本被表示成短语集合的情况下，其基本思想是采用两个文本之间相交的短语占两个文本短语并集的比例作为文本的相似度。
聚类算法		
基于层次的聚类算法：融合方法和分裂方法 基于划分的聚类算法：比较典型的是 K-means 算法 基于密度的聚类算法：比较典型的是 DBSCAN 算法 基于网格和子空间的聚类算法		

为高层级概念实现粗粒度的聚类，识别不同题材的文本，再结合各层级概念与非概念的特征词实现细粒度的聚类，揭示不同深度的主题信息。文献[43]提出将文本聚类与 LDA 相融合的微博主题检索模型，在对对应索引的频繁词集进行文本聚类后，调用每个类簇的 LDA 算法，从而挖掘出潜在的主题。除此之外，通过聚类进行主题识别也是主题演化最早且最常用的方法，主题演化的发展历史也是不断优化文本主题语义信息以提升聚类精度的过程，随着语义资源的丰富和语义标注技术的成熟，通过文本语义挖掘方法使得越来越多的语义信息在主题中得以体现，成为主题演化方法的一种重要发展趋势。

(2) 文本分类

文本分类是将自由文本文献自动归入一个或多个事先制定好的类目中，对文本进行有效的组织和管理，便于用户准确定位所需要的信息，是解决大数据环境下信息过载的关键技术之一。

文本分类主要包括文本表示、特征选择和分类器训练三个环节，在文本表示上文本分类与文本聚类的方法基本一致，不再赘述。表 3 详细列出了特征选择和分类器训练环节所使用的主要的方法和算法并进行了比较分析。

表 3 文本分类中特征选择和分类器训练环节方法与算法总结

分类环节	方法	比较分析
特征选择	文档频率(DF)	能够很容易地用于大规模语料统计。
	信息增益(IG)[44]	在机器学习领域被广泛使用。
	统计量(CHI)[45]	目前最好的特征选择方法之一，与其他方法相比，减少了约 50%的词汇，分类效果好，在文本数量逐减增多过程中，稳定性好，大多数的中文分类系统都采用该方法。
	互信息(MI)[46]	在统计语言模型中被广泛使用。
分类器训练	支持向量机(SVM)	中英文分类中分类精度最高，但是时间开销最大。
	KNN	K 近邻算法简单，易于实现，但是分类精确度不高。
	贝叶斯	所需训练时间最少，但是在特征项之间联系特别紧密的情况下，分类性能受到较大影响。
	决策树	产生的分类规则更易于理解且能够很容易地用于离散型的属性数据，但是当属性值较多时会受到影响。
	神经网络算法	不需要先验知识，但是内部规则的可理解性差，很难从中提取规则，面对离散型属性数据需要先转化成数值属性，在属性较多时受到的影响比决策树更大。

除了在网络信息处理上的广泛应用外[47-48]，文本分类方法在当前情报分析过程中的主流应用之一就是信息抽取中实体关系的识别，文献[49]将实体关系的识别看作分类问题，通过KNN、SVM等分类方法训练分类器，实现半监督的实体关系抽取任务。电子病历包含大量与患者健康状况密切相关的医疗知识，因此对它们的识别是信息抽取在医疗领域的重要进展。文献[50]梳理了命名实体识别和关系抽取的方法，分析了电子病历命名实体识别、实体修饰识别和实体关系抽取的主要方法，在此基础上总结出电子病历实体关系抽取主要采用机器学习的分类方法，并一般采用SVM和最大熵等。

3 文本语义挖掘方法在情报分析中的应用

科学研究领域的发展变化跟踪和研究前沿预测是情报分析方法的重要应用，尤其在当前创新驱动发展的形势下，为决策者提供数据支持是情报工作人员的重要职责。在激烈的科技竞争中能够准确及时地把握科学研究的发展脉络并进行研究主题的前瞻性预测是占据科技制高点的关键，科学数据是完成上述情报工作的最理想的数据源，当前情报分析在这方面的主要应用包括通过科学论文数据的主题演化分析和科学研究前沿探测以及通过专利数据的新兴技术挖掘。

3.1 主题演化分析

学科主题演化[42]是指以词语为表征的学科主题在时间维度上的发展变化过程，与空间变化相比，学科主题的时间演化体现的是学科主题的新陈代谢过程，体现某一学科的发展态势和未来走向，是研究学科发展规律的重要内容，文本语义挖掘方法在主题演化分析中主要体现在通过聚类进行主题识别，目前主要是引文分析和共词分析的方法。

引文分析的方法[51-52]是通过连续时间段内的共引聚类图的历时比较揭示科学研究主题的演化，由于其涉及的只是外部指标，没有使用含有主题信息最多的文本语义信息，在主题的确定上缺乏准确性，因此较之共词分析的方法有所不足。

(1) 共词分析法的发展

共词分析的基本原理是统计一组词两两在同一篇文献中出现的次数，以此为基础对这些词进行聚类分析，从而判定这些词之间的亲疏关系，分析这些词所代表的学科和主题结构的变化。1986年，Callon等[53]首次在饮食纤维等研究领域通过主题词共词分析，建立不同时期的网络图谱，分析了主题词在不同时期的网络图谱中的变化情况以揭示该领域的主题演化轨迹。1995年，国内崔雷[54]以丙型肝炎作为研究主题，选取1991年–1992年间与该主题相关的高被引文献，进行主题词的共词聚类分析，将1992年高被引论文进行同被引聚类并进行比较，对比揭示了丙型肝炎领域研究主题的变化情况，并进行情报预测，是基于主题词的共词分析方法在主题演化以及研究前沿探测的较早研究。后来国内外学者[55-56]又基于文本的关键词、主题词或高频词直接进行共词聚类分析，揭示了不同领域的学科主题演化轨迹。

但是共词分析的方法都是基于词频的聚类，在文本主题的呈现上不够充分，因此有学者开始通过文本语义挖掘的方法改善文本主题表示，提升了主题演化的准确性和精确性。

①优化关键词的选择

王晓光[57]研究证明共词网络内存在社区现象，通过层次聚类识别网络社区，这些社区由许多节点组成，每一个节点都是文章的关键词，因此在语义上大大丰富了主题表示，并通过社区相似度算法构建了科研主题演化分析模型来发现研究前沿。文献[58]提出一种基于K-clique社区的知识创新演化揭示方法，首先将2008年–2012时间段内碳纳米管领域的文献构造时序关键词共词网络，利用C-Finder生成K-clique社区，使用Sybase公司的PowerBuilder进行演化处理，准确识别了碳纳米管领域该时间段内的知识创新主要方向。文献[59]将TF-IDF引入筛选重要的关键词，并用滑动时间窗口切分数据，构建共词网络，抽取网络的最大连接图进行聚类，并计算相似度从而完成研究主题的演化图谱，在LED领域从演化视角揭示了研究前沿，并揭示了研究前沿主题产生、成长、消退、消失的过程。

②优化语义关系

文献[60]针对共词分析法中忽略了关键词对之间的深层语义关系的问题，提出一种基于关键词共现和语义关联相结合的主题演化方法。通过Word2Vec将关键词表示成语义级别的词向量，并通过Pearson系数计算关键词之间的相关系数，从而准确识别出信息检索领域的主题演化趋势。文献[61]针对主题演化中的语义缺失和批处理问题，通过语义角色标注技术改善文本主题表示，提出一种在线增量的基于特征本体的主题演化方法，首先使用TF-IDF方法计算每个词对文档的重要程度，按重要程度将前N个词对应的词语–文档矩阵转化为词语–词语连接矩阵，保留高于某固定阈值的所有值，也就是词语之间的关系值，然后对词语之间的

语义关系进行包括同义词(synonym)、上义词(hypernym)、反义词(antonym)、整体(holonym)、蕴含(entailment)、致使(cause)、属性(attributeOf)、属性值(attribute)、实例(instance)、一般关系(relation) 10 种语义角色的标注，增强了语义的可解释性，形成由多个连通图组成的特征本体，每个连通图代表一个主题，大大丰富了文本的主题语义信息。国内学者叶春蕾等[42]提出一种利用改进的共词分析方法，在进行主题词抽取和共现矩阵构建后，利用 LDA 主题模型进行主题网络识别，该方法更能精确地体现主题词、主题和文档之间的三层语义关系，以信息量取代传统共词分析方法中以词频或共现词频作为主题识别的指标，更客观、准确地揭示主题的演化规律。

(2) 文本主题挖掘的发展

文献[62]基于层次概率模型hLDA并考虑时间信息自动挖掘科技文献中潜在的主题信息，利用Gibbs抽样方法对模型参数进行推断，同时利用互信息对主题词进行筛选，最终使用先/后离散分析方法研究科技文献的主题随时间的演化。文献[63]运用隐马尔可夫模型理论，选择网民特征、信息主题和信息内容完整度三维指标，设定隐马尔可夫的状态值，并选取舆情形成、发展、波动、消亡的观测值，构建了面向网络舆情发现的隐马尔可夫模型，揭示微博舆情的演化过程。文献[64]提出一种利用会议数据进行动态主题演化分析的方法，首先利用马尔可夫条件随机场对数据进行主题聚类，经过这一步可以对文本主题进行一个浅层的语义理解，而且不需要对聚类的主题数进行规定，利用MeSH与主题词之间进行映射，并选择合适的主题词作为聚类主题的标签，通过计算主题相似度实现了DBLP领域的主题演化分析。

本文将文本语义挖掘方法在主题演化分析中的应用进行总结，如图1所示:

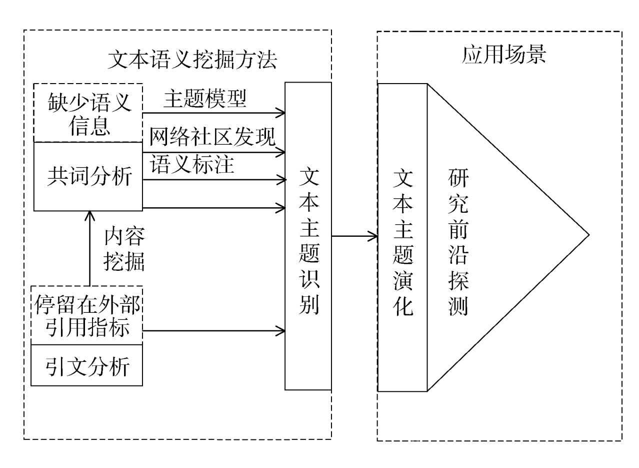


图 1 文本语义挖掘在主题演化分析中的应用

可以看出针对引文分析存在的弊端，文本语义挖掘方法以其深入到文本内容的分析呈现出特有的优越性，但是共词分析法也有一定的语义缺陷性，因此文本语义挖掘分别通过词粒度的标注和聚类算法对共词分析法进行改善，除此之外还有通过其他主题模型进行文本主题优化表示的方法。本文所介绍的各粒度的文本挖掘方法在主题识别阶段都得到广泛的应用，而且成为针对传统情报分析方法的语义缺失问题进行改善的重要手段，由于其处理粒度的灵活性以及丰富性，文本语义挖掘方法既可以作为语义补充参与其他方法进行主题演化分析，也可以形成完整的主题演化分析流程，有着广阔的应用前景。

3.2 技术挖掘

技术挖掘[65]是 21 世纪初，美国学者 Porter 等提出的基于历史科技文献分析当前和未来技术发展现状和趋势的理论和方法。与其他的科技文献相比，专利具有结构规范，技术叙述详尽、严谨以及分类科学等特点，更容易表示成结构化的语义模型，因此成为技术挖掘使用最多的信息源。专利的文本挖掘是文本挖掘在专利文献中的应用，其核心部分在专利文本知识表示和专利技术的演化趋势分析。

目前专利文本知识表示的主流方法是包含语义信息的向量空间知识表示，其中 SAO (Subject-Action-Object)语义向量又是在专利文本中最常使用的方法，该方法通过语义标注技术、命名实体识别技术和文本分类技术实现专利 SAO 结构的抽取。相对于基于本体知识、技术向量空间模型的专利文本知识表示方法，SAO 结构的专利文本表示方法既省时省力，又有语义信息的补充，为进一步的技术挖掘提供了良好的基础。

SAO 结构源自于发明问题解决理论，面向专利的 SAO 结构抽取是从文本中抽取出(Subject, Action, Object)实体关系三元组，是表示问题解决方法的基本功能函数单元，在专利文献中能够呈现各概念之间的关系，可以将专利的核心部分表示出来。S 和 O 代表部件实体，一般由名词或名词性短语表示，通过词性标注和进一步的命名实体识别进行抽取。S 和 O 之间的相互关系的指示词 A 的抽取则是难点，属于实体关系抽取，通过分类方法采用机器学习的方法完成。

Yoon 等[66]为了挖掘碳纳米管领域的最新技术，在对该领域专利文献进行句子切分、词性标注、词义消

歧等预处理后，抽取出 SAO 结构，如表 4 所示：

表 4　SAO 结构抽取结果(Patent EP02749069)

S (subject)	A (action)	O (object)
Carbon	Contain	Gas plasma
Chemical vapor deposition	Use	Carbon
Group	Comprise	Nitrogen, hydrogen argon and ammonia
Microwave energy	Generate	Plasma
Plasma chamber	Cool	Electrodes
Powder	Have	Particle size
CVD chamber	Inject	Catalyst
Vacuum chamber	Generate	Gaseous plasma
Vacuum chamber	Maintain	Gaseous plasma

表 4 中的第一行 Carbon 和 Gas plasma 都是名词性质，通过词性标注加以识别抽取，Contain 是它们之间的关系，通过语义标注来识别并通过基于机器学习的分类来进行大规模抽取，即“碳含有气体等离子”。将所有的专利文本表示成相应的 SAO 结构之后，专利之间的相似度对比就变成了专利所含的 SAO 结构的句子之间的文本语义相似度计算，得到专利的相似度矩阵，并形成专利网络图，通过定义 DSI 与 GCI 两个指标来识别先进技术。文献[67]为了通过基于功能的专利分析来挖掘技术潜在应用领域以便支持技术转移，构造专利的 SAO 结构中 AO 部分的语义向量空间，在 WordNet 的帮助下通过比较 Action 的相似度，识别具有定义功能的专利，将这些专利与工业领域类别进行映射，从而识别出某技术的工业应用领域。文献[68]在对专利进行 SAO 结构表示的基础上，将专利进行相似度计算，并进行多维尺度分析，找到离群点，这些离群点有些即代表了新兴技术。以上研究在进行结构抽取时都使用了自然语言的处理工具，目前这些工具都已经具有词性标注、句法分析等功能，完全可以辅助实现结构化信息抽取。

胡正银等[69]在对专利进行 SAO 结构进行抽取的基础上，对每一条 SAO 按照技术问题、技术方案、技术功能与技术效果再进行语义标注，如 Action 为“used as”等，则该 SAO 被标注为“Function”语义类型；如果 Action 为“comprise”，则该 SAO 被标注为“Solution”语义类型。在对 SAO 结构进行语义标注后，进行降维并生成技术主题，并通过文本聚类算法进行技术主题聚类，生成技术主题演化图，识别新的技术方向。

本文将文本语义挖掘方法在专利技术挖掘中的应用进行总结，如图 2 所示。可以看出，当前文本语义挖掘方法在专利技术挖掘的主流应用集中在通过文本语义挖掘方法对专利文本的表示，并通过文本相似度算法进行技术演化和新兴技术的识别。词语和句子粒度的语义挖掘方法已经成为结构化信息抽取的基础，而且该方法已经比较成熟，今后在技术挖掘方面的研究会更多地集中在新兴技术主题的识别和判定上。

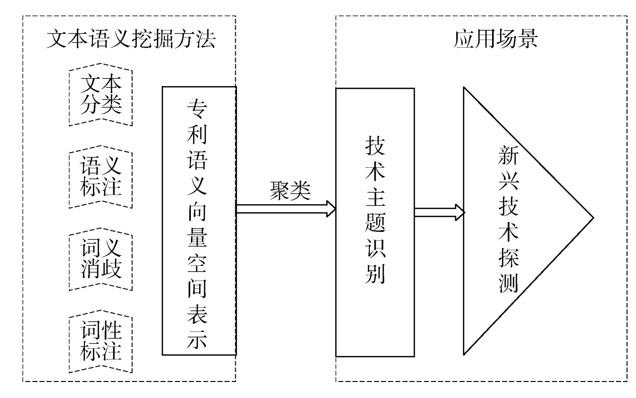


图 2　文本语义挖掘方法在技术挖掘中的应用

4　结　语

针对传统的情报分析方法侧重于分析结构化数据的局限性，本文从词、句子和篇章粒度分别介绍了当前的主要文本语义挖掘方法，该方法以其处理粒度的灵活性以及对文本格式的普适性，完全能够应对情报分析工作中遇到的格式问题；针对情报分析中的语义缺失问题，当前的文本语义挖掘方法实现了词语、句子和篇章粒度的语义丰富，充分挖掘信息中蕴含的情报，大大提高了情报分析的准确性，使得文本语义挖掘方法在情报工作中有着广阔的应用前景。

从当前的发展情况来看，多种异构的数据源已经不是情报工作中的处理难题，下一步的主要研究重点在于：文本语义挖掘的自动化，当前的文本语义挖掘方法大都是有监督或半监督的方法，精确度较高，在数据量急剧扩张的情况下，无监督的文本语义挖掘方法可以有效提高情报工作效率；实现文本语义资源的进一步融合，目前通过外部语义资源进行语义丰富辅助文本挖掘的方法还不是很成熟，在各种语义挖掘技术出现并成熟的情况下，例如本体、语义词典、语义网等语义资源的构建将大大影响到文本语义挖掘的效果。

参考文献：

[1] Kantardzic M. 数据挖掘：概念、模型、方法和算法[M]. 王晓海，吴志刚译. 北京：清华大学出版社，2013：250-251. (Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms [M]. Translated by Wang Xiaohai, Wu Zhigang. Beijing: Tsinghua University Press, 2013: 250-251.)

[2] 王丽杰，车万翔，刘挺. 基于 SVMTool 的中文词性标注[J]. 中文信息学报，2009，23(4)：16-21. (Wang Lijie, Che Wanxiang, Liu Ting. An SVMTool-Based Chinese POS Tagger [J]. Journal of Chinese Information Processing, 2009, 23(4): 16-21.)

[3] 张民，李生，赵铁军，等. 统计与规则并举的汉语词性自动标注算法[J]. 软件学报, 1998, 9(2): 134-138. (Zhang Min, Li Sheng, Zhao Tiejun, et al. Part of Speech Tagging Chinese Corpus Based on Statistics and Rules[J]. Journal of Software, 1998, 9(2): 134-138.)

[4] 郭永辉，吴保民，王炳锡. 一种用于词性标注的相关投票融合策略[J]. 中文信息学报，2007，21(2)：9-13. (Guo Yonghui, Wu Baomin, Wang Bingxi. Correlation Voting Fusion Strategy Used for Part of Speech Tagging [J]. Journal of Chinese Information Processing, 2007, 21(2): 9-13.)

[5] 洪铭材，张阔，唐杰，等. 基于条件随机场 CRFs 的中文词性标注方法[J]. 计算机科学, 2006, 33(10): 148-155. (Hong Mingcai, Zhang Kuo, Tang Jie, et al. A Chinese Part-of-Speech Tagging Approach Using Conditional Random Fields[J]. Computer Science, 2006, 33(10): 148-155.)

[6] 张民，李生，赵铁军，等. 统计与规则并举的汉语词性自动标注算法[J]. 软件学报, 1998, 9(2): 134-138. (Zhang Min, Li Sheng, Zhao Tiejun, et al. Part of Speech Tagging Chinese Corpus Based on Statistics and Rules[J]. Journal of Software, 1998, 9(2): 134-138.)

[7] ICTCLAS[K]. [2015-07-28]. http://ictclas.nlpir.org/.

[8] 哈工大语言云[K]. [2015-08-13]. http://www.ltp-cloud.com/. (LTP[K]. [2015-08-13]. http://www.ltp-cloud.com/.)

[9] Stanford Log-linear Part-Of-SpeechTagger[K]. [2015-09-15]. http://nlp.stanford.edu/software/tagger.shtml.

[10] CLAWS POS Tagger[K]. [2015-09-18]. http://ucrel.lancs.ac.uk/claws/trial.html.

[11] NLTK [K]. [2015-07-20]. http://www.nltk.org/.

[12] 商宪丽，王学东.微博话题识别中基于动态共词网络的文本特征提取方法[J]. 图书情报知识，2016(3)：80-88. (Shang Xianli, Wang Xuedong. A Feature Selection Method Based on Dynamic Co-word Network for Microblog Topic Detection [J]. Documentation, Information & Knowledge, 2016(3): 80-88.)

[13] 杜思奇，李红莲，吕学强. 基于汉语组块分析的情感标签抽取[J]. 情报理论与实践，2016，39(5)：125-129. (Du Siqi, Li Honglian, Lv Xueqiang. Chinese Chunking Based Emotional Label Extraction [J]. Information Studies: Theory & Application, 2016, 39(5): 125-129.)

[14] 兰秋军，刘文星，李卫康，等. 融合句法信息的金融论坛文本情感计算研究[J]. 现代图书情报技术，2016(4)：64-71. (Lan Qiujun, Liu Wenxing, Li Weikang, et al. Sentiment Analysis of Financial Forum Textual Message [J]. New Technology of Library and Information Service, 2016(4): 64-71.)

[15] 翟羽佳，王芳. 基于文本挖掘的中文领域本体构建方法研究[J]. 情报科学, 2015, 33(6): 3-10. (Zhai Yujia, Wang Fang. Research on Construction Methods of Chinese Domain Ontology Based on Text Mining [J]. Information Science, 2015, 33(6): 3-10.)

[16] 吴云芳. 词义消歧研究：资源、方法与评测[J]. 当代语言学, 2009, 11(2): 113-123. (Wu Yunfang. A Survey of Chinese Word Sensedisambiguation: Resources, Methods and Evaluation [J]. Contemporary Linguistics, 2009, 11(2): 113-123.)

[17] 卢志茂，刘挺，李生. 统计词义消歧的研究进展[J]. 电子学报, 2006, 34(2): 333-343. (Lu Zhimao, Liu Ting, Li Sheng. The Research Progress of Statistical Word Sense Disambugation [J]. Electronic Sinica, 2006, 34(2): 333-343.)

[18] Lesk M E. Automated Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from All Ice Cream Cone[C]. In: Proceedings of the S1GDOC Conference. New York: Association for Computing Machinery, 1986: 24-26.

[19] Pook S L, Catlett J. Making Sense out of Searching[R]. Sydney: AT&T Bell Laboratories, 1988.

[20] Agirre E, Rigau G. A Proposal for Word Sense Disambiguation Using Conceptual Distance [C]. In: Proceedings of the 1st International Conference on Recent Advances in NLP. 1995: 162-171.

[21] 鹿文鹏，黄河燕，吴昊. 基于领域知识的图模型词义消歧方法[J]. 自动化学报，2014，40(12)：2836-2850. (Lu Wenpeng, Huang Heyan, Wu Hao. Word Sense Disambiguation Based with Graph Model Based on Domain Knowledge [J]. Acta Automatic Sinica, 2014, 40(12): 2836-2850.)

[22] 张仰森，郭江. 四种统计词义消歧模型的分析与比较[J]. 北京信息科技大学学报，2011，26(2)：13-18. (Zhang Yangsen, Guo Jiang. Analysis and Comparison of 4 Kinds of Statistical Word Sense Disambiguation Models[J]. Journal of Beijing Information Science & Technology, 2011, 26(2):

13-18.)

[23] 鲁松, 白硕, 黄雄, 等. 基于向量空间模型的有导消歧[J]. 计算机研究与发展, 2011, 38(6): 662-667. (Lu Song, Bai Shuo, Huang Xiong, et al. Supervised Word Sense Disambiguation Based on Vector Space Model [J]. Computer Research and Development, 2011, 38(6): 662-667.)

[24] 王瑞琴, 孔繁胜. 无监督词义消歧研究[J]. 软件学报, 2009, 20(8): 2138-2152. (Wang Ruiqin, Kong Fansheng. Unsupervised Word Sense Disambiguation Research [J]. Journal of Software, 2009, 20(8): 2138-2152.)

[25] BRAT [K]. [2015-09-18]. http://brat.nlplab.org/index.html.

[26] 杨建林, 王文龙. 公共卫生类突发事件的抽取研究[J]. 情报理论与实践, 2016, 39(4): 51-59. (Yang Jianlin, Wang Wenlong. Public Sanitation Emergency Event Extraction[J]. Information Studies: Theory & Application, 2016, 39(4): 51-59.)

[27] 陈锋, 翟羽佳, 王芳. 基于条件随机场的学术期刊中理论的自动识别方法[J]. 图书情报工作, 2016, 60(2): 122-128. (Chen Feng, Zhai Yujia, Wang Fang. Automatic Theory Recognition in Academic Journals Based on CRF [J]. Library and Information Service, 2016, 60(2): 122-128.)

[28] 祝娜, 王效岳, 白如江. 语义角色标注及其在科技情报分析中的应用研究[J]. 情报理论与实践, 2015, 38(1): 98-103. (Zhu Na, Wang Xiaoyue, Bai Rujiang. Semantic Role Labeling and the Application in Intelligence Analysis [J]. Information Studies: Theory & Application, 2015, 38(1): 98-103.)

[29] Hacioglu K. Semantic Role Labeling Using Dependency Trees [C]. In: Proceedings of the 20th International Conference on Computational Linguistics. Association for Computational Linguistics, 2004.

[30] 王步康, 王红玲, 袁晓虹, 等. 基于依存句法分析的中文语义角色标注[J]. 中文信息学报, 2010, 24(1): 25-29. (Wang Bukang, Wang Hongling, Yuan Xiaohong, et al. Chinese Dependency Parse Based Semantic Role Labeling [J]. Journal of Chinese Information Processing, 2010, 24(1): 25-29.)

[31] Gildea D, Palmer M. The Necessity of Parsing for Predicate Argument Recognition [C]. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002: 239-246.

[32] Pradhan S, Ward W, Hacioglu K, et al. Shallow Semantic Parsing Using Support Vector Machines [C]. In: Proceedings of HLT-NAACL.2004: 233-240.

[33] 李世奇, 赵铁军, 李晗静, 等. 基于特征组合的中文语义角色标注[J]. 软件学报, 2011, 22(2): 222-232. (Li Shiqi, Zhao Tiejun, Li Hanjing, et al. Chinese Semantic Role Labeling Based on Feature Combination [J]. Journal of Software, 2011, 22(2): 222-232.)

[34] 王红玲. 基于特征向量的中英文语义角色标注研究[D]. 苏州: 苏州大学, 2009. (Wang Hongling. Chinese and English Semantic Role Labeling Based on Feature Vector [D]. Suzhou: Soochow University, 2009.)

[35] 宋毅君, 王瑞波, 李济洪, 等. 基于条件随机场的汉语框架语义角色自动标注[J]. 中文信息学报, 2014, 28(3): 36-47. (Song Yijun, Wang Ruibo, Li Jihong. et al. Semantic Role Labeling of Chinese FrameNet Based on Conditional Random Fields [J]. Journal of Chinese Information Processing, 2014, 28(3): 36-47.)

[36] 李明, 王亚斌, 张其文, 等. 基于树状条件随机场模型的语义角色标注[J]. 计算机工程, 2010, 36(18): 41-45. (Li Ming, Wang Yabin, Zhang Qiwen, et al. Semantic Role Labeling Based on Tree Conditional Random Fields Model[J]. Computer Engineering, 2010, 36(18): 41-45.)

[37] 白如江, 祝娜, 王效岳. 语义增强的科技创新内容表征研究[J]. 情报理论与实践, 2016, 39(3): 73-79. (Bai Rujiang, Zhu Na, Wang Xiaoyue. Semantic Representation of Technical Innovation Content Based on Semantic Enhancement [J]. Information Studies: Theory & Application, 2016, 39(3): 73-79.)

[38] 张帆, 乐小虬. 领域科技文献创新点句中主题属性实例识别方法研究[J]. 现代图书情报技术, 2015(5): 15-23. (Zhang Fan, Le Xiaoqiu. Research on Recognition of Concept Attribute Instances in Innovation Sentences of Scientific Research Paper [J]. New Technology of Library and Information Service, 2015 (5): 15-23.)

[39] 祝娜, 王效岳, 杨京, 等. 基于 LDA 的科技创新主题语义识别研究[J]. 图书情报工作, 2015, 59(14): 126-134. (Zhu Na, Wang Xiaoyue, Yang Jing, et al. Semantic Recognition of Technological Innovation Theme Based on LDA[J]. Library and Information Service, 2015, 59 (14): 126-134.)

[40] 洪韵佳, 许鑫. 基于领域本体的知识库多层次文本聚类研究——以中华烹饪文化知识库为例[J]. 现代图书情报技术, 2013(12): 19-26. (Hong Yunjia, Xu Xin. Study on Multi-Level Text Clustering for Knowledge Base Based on Domain Ontology——Taking Knowledge Base of Chinese Cuisine Culture as an Example [J]. New Technology of Library and Information Service, 2013(12): 19-26.)

[41] 常娥. 基于 LSI 理论的文本自动聚类研究[J]. 图书情报工作, 2012, 56(11): 89-92. (Chang E. Automatic Text Clustering Based on Latent Semantic Index Theory [J]. Library and Information Service, 2012, 56(11): 89-92.)

[42] 叶春蕾, 冷伏海. 基于共词分析的学科主题演化方法改进

研究[J]. 情报理论与实践, 2012, 35(3): 79-82. (Ye Chunlei, Leng Fuhai. Development of Discipline Theme Evolution Analysis Based on Co-word Analysis [J]. Information Studies: Theory & Application, 2012, 35(3): 79-82.)

[43] 唐晓波, 房小可. 基于文本聚类与 LDA 相融合的微博主题检索模型研究[J]. 情报理论与实践, 2013, 36(8): 85-90. (Tang Xiaobo, Fang Xiaoke. Micro Blog Topic Retrieval Model Research Based on Text Clustering and LDA[J]. Information Studies: Theory & Application, 2013, 36(8): 85-90.)

[44] Mitchell T. Machine Learning [M]. McCraw Hill, 1996.

[45] Yang Y. An Evaluation of Statistical Approaches to Text Categorization [J]. Information Retrieval, 1999, 1(1-2): 69-90.

[46] Church K W, Hanks P. Word Association Norms, Mutual Information and Lexicography[J]. Computational Linguistics, 1990, 16(1): 22-29.

[47] Google 新闻的工作原理[EB/OL]. [2016-04-28]. http://support, google. com/news/ bin/ topic. py? Hl=zh-Hans& topic =2428790. (The Working Principle of Google News [EB/OL]. [2016-04-28].http://support, google. com/news/ bin/ topic. py? Hl=zh-Hans&topic =2428790.)

[48] 新华网[EB/OL]. [2016-04-28]. http://baike. baidu.com/view/ 154954. htm. (xinhuanet [EB/OL]. [2016-04-28]. http://baike. baidu. com/view/154954. htm.)

[49] 宁海燕. 实体关系自动抽取技术的比较研究[D]. 哈尔滨: 哈尔滨工业大学, 2010. (Ning Haiyan. Comparative Study of Automatic Entity Relation Extraction [D]. Harbin: Harbin Insititute of Technology, 2010.)

[50] 杨锦锋, 于秋滨, 关毅, 等. 电子病历命名实体识别和实体关系抽取研究综述[J]. 自动化学报, 2014, 40(8): 1537-1560. (Yang Jinfeng, Yu Qiubin, Guan Yi, et al. An Overview of Research on Electronic Medical Record Oriented Named Entity Recognition and Entity Relation Extraction[J]. Acta Automatic Sinica, 2014, 40(8): 1537-1560.)

[51] 侯跃芳, 崔雷, 吴迪. 应用引文共引聚类–内容词分析法对学科发展的研究[J]. 情报学报, 2007, 26(2): 309-314. (Hou Yuefang, Cui Lei, Wu Di. Co-Citation Clustering-Content Words Analysis in Subject Development [J]. Journal of the China Society for Scientific and Technical Information, 2007, 26(2): 309-314.)

[52] 柴省三. 内容词–共引聚类分析及其在科学结构研究中的应用[J]. 情报学报, 1997, 16(1): 68-73. (Chai Shengsan. Application of Content Words and Co-citation Clustering Analysis to Science Structure Studies[J]. Journal of the China Society for Scientific and Technical Information, 1997, 16(1): 68-73.)

[53] Callon M, Law J, Rip A. Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World[M]. London: The Macmillan Press LTD, 1998.

[54] 崔雷. 当年高被引论文的主题词链聚类分析及其在情报预测中的应用[J]. 情报学报, 1995, 14(5): 368-373. (Cui Lei. Keyword Link Cluster Analysis of the Immediately Highly Cited Papers and Its Utilization in Information Prediction [J]. Journal of the China Society for Scientific and Technical Information, 1995, 14(5): 368-373.)

[55] Callon M, Courtial J P, Laville F. Co-word Analysis as a Tool for Describing the Network of Interactions Between Basic and Technological Research: The Case of Polymer Chemistry [J]. Scientometrics, 1991, 22(1): 155-205.

[56] Kostoff R N, Eberhart H J, Toothman D R. Data-base Tomography for Technical Intelligence: A Roadmap of The Near-earth Space Science and Technology Literature [J]. Information Processing & Management, 1997, 34(1): 69-85.

[57] 王晓光. 科学知识网络的结构与演化(I): 共词网络方法的提出[J]. 情报学报, 2009, 28(4): 599-605. (Wang Xiaoguang. Structure and Evolution of Scientific Knowledge Network: Co-word Network[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(4): 599-605.)

[58] 白如江, 冷伏海. k-clique 社区知识创新演化方法研究[J]. 图书情报工作, 2013, 57(17): 86-94. (Bai Rujiang, Leng Fuhai. Knowledge Innovational Evolution Analysis Based on k-clique Community Network [J]. Library and Information Service, 2013, 57(17): 86-94.)

[59] 郑彦宁, 许晓阳, 刘志辉. 基于关键词共现的研究前沿识别方法研究[J]. 图书情报工作, 2016, 60(4): 85-92. (Zheng Yanning, Xu Xiaoyang, Liu Zhihui. Study on the Method of Identifying Research Fronts Based on Keywords Co-occurrence [J]. Library and Information Service, 2016, 60(4): 85-92.)

[60] 巴志超, 杨子江, 朱世伟, 等. 基于关键词语义网络的领域主题演化分析方法研究[J]. 情报理论与实践, 2016, 39(3): 67-72. (Ba Zhichao, Yang Zijiang, Zhu Shiwei, et al. Key Words Semantic Network Based Field Topic Evolution Analysis Model [J]. Information Studies: Theory & Application, 2016, 39(3): 67-72.)

[61] 陈千, 桂志国, 郭鑫, 等. 基于特征本体的文本流主题演化[J]. 计算机应用, 2015, 35(2): 456-460. (Chen Qian, Gui Zhiguo, Guo Xin, et al. Topic Evolution in Text Stream Based on Feature Ontology [J]. Journal of Computer Applications, 2015, 35(2): 456-460.)

[62] 王平. 基于层次概率主题模型的科技文献主题发现及演化

[J]. 图书情报工作, 2014, 58(22): 70-77. (Wang Ping. Topic Extraction and Evolution for Scientific Literature Based on Hierarchical Probabilistic Topic Model [J]. Library and Information Service, 2014, 58(22): 70-77.)

[63] 何建民, 李雪. 面向微博舆情演化分析的隐马尔科夫模型研究[J]. 情报科学, 2016, 34(4): 7-12. (He Jianmin, Li Xue. A Hidden Markov Model Research in the Microblog Public Opinion Evolutionary Analysis [J]. Information Science, 2016, 34(4): 7-12.)

[64] Song M, Heo G E, Kim S Y. Analyzing Topic Evolution in Bioinformatics: Investigation of Dynamics of the Field with Conference Data in DBLP[J]. Scientometrics, 2014, 101(1): 397-428.

[65] 胡正银, 方曙. 专利文本技术挖掘研究进展综述[J]. 现代图书情报技术, 2014(6): 62-70. (Hu Zhengyin, Fang Shu. Review of Patent Text Technology Mining Research Development [J]. New Technology of Library and Information Service, 2014(6): 62-70.)

[66] Yoon J, Kim K. Identifying Rapidly Evolving Technological Trends for R&D Planning Using SAO-based Semantic Patent Networks [J]. Scientometrics, 2011, 88(1): 213-228.

[67] Park H, Yoon J, Kim K. Using Function-based Patent Analysis to Identify Potential Application Areas of Technology for Technology Transfer [J]. Expert Systems with Applications, 2013, 40(13): 5260-5265.

[68] Yoon J, Kim K. Detecting Signals of New Technological Opportunities Using Semantic Patent Analysis and Outlier Detection [J]. Scientometrics, 2012, 90(2): 1-17.

[69] 胡正银, 方曙, 隗玲. 基于 SAO 的专利技术演化分析[C]. 见: 中国图书馆学会专业图书馆分会 2015 年年会论文集, 贵阳. 2015. (Hu Zhengyin, Fang Shu, Kui Ling. Patent Technology Evolution Analysis Based on SAO [C]. In: Proceedings of Professional Library Branch of China Library Association 2015 Scholar Conference, Guiyang. 2015.)

作者贡献声明:

赵冬晓: 论文撰写, 资料收集和调研;
王效岳: 论文最终版本修订;
白如江: 提出论文研究思路;
刘自强: 论文修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 赵冬晓. article.zip. 涉及文中总结的文本语义挖掘方法的论文资料.

[2] 赵冬晓. graph. zip. 有助于文中结论分析的图表数据.

收稿日期: 2016-06-06
收修改稿日期: 2016-07-21

Semantic Text Mining Methodologies for Intelligence Analysis

Zhao Dongxiao Wang Xiaoyue Bai Rujiang Liu Ziqiang
(Institute of Scientific & Technical Information, Shandong University of Technology, Zibo 255049, China)

Abstract: [**Objective**] This paper reviews the semantic text mining techniques for intelligence analysis. [**Coverage**] We surveyed the leading semantic text mining research on intelligence analysis from the last ten years and a few earlier studies. [**Methods**] We first discussed the semantic text mining methodologies and algorithms for words, sentences and paragraphs. Then, we analyzed these techniques from the perspective of topic evolution and applications of mining technologies. [**Results**] Compared to the traditional intelligence analysis methods, semantic text mining approaches could process unstructured data and deal with multi-layer structured data. [**Limitations**] Only reviewed the leading studies and their applications in the scientific field. [**Conclusions**] Semantic text mining improve the performance of traditional intelligence analysis systems and become the future direction of research methodology. More research is needed to enrich the outlier semantic resources.

Keywords: Semantic text mining Intelligence analysis Topic evolution Technology mining